TROIS NOUVEAUX INDICES DE RÉALISME DANS L'AUTO-ÉVALUATION DES PERFORMANCES

Dieudonné Leclercq & Marianne Poumay
Université de Liège - Belgique

1. LA CONNAISSANCE PARTIELLE, SA MESURE ET L'INTÉRÊT DE LA MÉTACOGNITION

Le philosophe anglais Bertrand Russel disait : « Le problème, dans notre monde, est que les imbéciles sont sûrs de tout et les sages pleins de doutes » et l'écrivain américain Mark Twain : « Ce n'est pas ce que nous ignorons qui nous nuit. C'est ce que dont nous sommes sûrs, mais qui est faux». Nous pensons que la connaissance du degré de confiance que l'on peut avoir dans ses propres connaissances fait partie de la connaissance d'une personne et devrait être évalué en tant que tel. En fait, nous nous rallions aux deux volets de la thèse de Bruno deFinetti (1965) « La connaissance partielle existe. La détecter est nécessaire et faisable » (p. 109) et « Seule la probabilité subjective peut donner une signification objective à toute réponse et toute méthode de notation. » (p. 111).

2. UNE LABORIEUSE RECHERCHE DES CONSIGNES ET PROCÉDURES OPTIMALES

Il a fallu beaucoup de temps et d'errements à la communauté scientifique pour découvrir quelles méthodes de recueil, de traitement, de scoring, d'interprétation étaient erronées, inefficaces, comportant des effets pervers, etc.

Ainsi, comme nous l'avons montré (Leclercq, 1983), les recherches des années 70 aux Etats-Unis et aux Pays Bas manquent de validité théorique par deux aspects. D'une part par leur recours à des consignes ordinales du type « peu sûr, moyennement sûr, extrêmement sûr », ou « met Z ou zonder Z » (Z signifiant Zekerheid) (Sandbergen, 1971) ou encore « 1. I guess, 2. Fairly sure, 3. Confident » (Jacobs, 1971), consignes évidemment imprécises et donc ambiguës, débouchant sur des données ininterprétables. D'autre part par l'usage de barèmes de tarifs non inspirés de la théorie des décisions tels que « +1, +2,+3 » pour les Réponses Correctes (RC) avec les degrés de certitude 1, 2 et 3 et « 0, -2, -3 » pour les Réponses Incorrectes (RI) avec ces degrés de certitude

(Jacobs, 1972). Des tentatives d'améliorations sont venues de consignes précisant des intervalles sur l'échelle des probabilités (ou des pourcentages de chances) telles que « 0= entre 0% et 25%, 1 = entre 25 et 50%, 2 = entre 50 et 75%, 3 = entre 75 et 100% » , et de barèmes de tarifs conformes à la théorie des décisions tels que « 0, +3, +4, +5 » pour les RC et « 0, -1, -2, -5 » pour les RI. (Leclercq, 1983, 198).

Ces consignes se sont raffinées par la mise en évidence des limites de la capacité humaine à discriminer de façon fiable des degrés différents sur l'échelle des probabilités. Leclercq (1983, 241-255) a en effet montré que, comme G. Miller (1956) l'avait montré pour d'autres domaines, cette limite était d'environ 7 portions différentes, pas plus et que, comme Stevens (1967) l'avait fait en psychophysique, la sensibilité était logarithmique, c'est-à-dire plus élevée aux deux extrémités (proches de 0 ou de 1). Malheureusement, ce raffinement de la consigne (6 à 7 portions de l'axe des probabilités, inégales entre elles) a accru la complexité tant dans les zones à prendre en considération que dans les calculs à effectuer, les tarifs devenant complexes et nécessitant une machine à calculer, voire un ordinateur. En 1999, un premier pas vers la simplification a été fait en adoptant une consigne simplifiée « Fournissez comme Degré de Certitude un des 6 pourcentages de chances suivants : 0%, 20%, 40%, 60%, 80%, 100%. » (Leclercq, 2003).

Le barème des tarifs, lui, restait jusqu'ici problématique lorsque les évaluations sont sanctionnantes, car, comme le montre Damasio (1994), les étudiants alors, pour maximiser leurs scores, ne se conforment pas à la théorie des décisions, mais adoptent des stratégies (comme fournir un même degré de certitude, 60% par exemple, pour toutes leurs réponses) incompatibles avec la mesure de la connaissance partielle. Nous examinerons d'abord les résultats engrangeables lorsque l'évaluation n'est pas sanctionnante, tout spécialement la **valeur diagnostique** des indices fournis aux étudiants sur leur métacognition, puis nous nous pencherons sur la situation d'évaluation sanctionnante où il importe que la métacognition soit exprimée en tant que telle par les étudiants, donc mesurées le plus validement possible et que son poids dans le score total à une épreuve soit ressenti comme **juste,** comme **pertinente** tant par les étudiants que par les enseignants. La combinaison des deux qualités (diagnosticité et pertinence) devant assurer la validité conséquentielle de ce type d'évaluation : sa capacité d'entraîner une réflexion métacognitive et des régulations en conséquence des stratégies de réponse et d'apprentissage.

3. UNE VASTE ÉVALUATION NON SANCTIONNANTE

Lors de l'opération MOHICAN[1] , environ 4000 étudiants ont été soumis, pour certains (des Sciences Humaines) à 6 épreuves et pour d'autres (des sciences de la nature) à 8 épreuves, le total étant de 123 questions chacun (Leclercq, 2003). Les questions étaient des QCM à 5 solutions proposées, plus deux solutions générales : Aucune et Toutes. En plus de la réponse à chaque question, les étudiants étaient invités à fournir leur degré de certitude (dans l'exactitude de la réponse) en choisissant un des 6 points de l'échelle suivante : 0%, 20%, 40%, 60%, 80%, 100%.

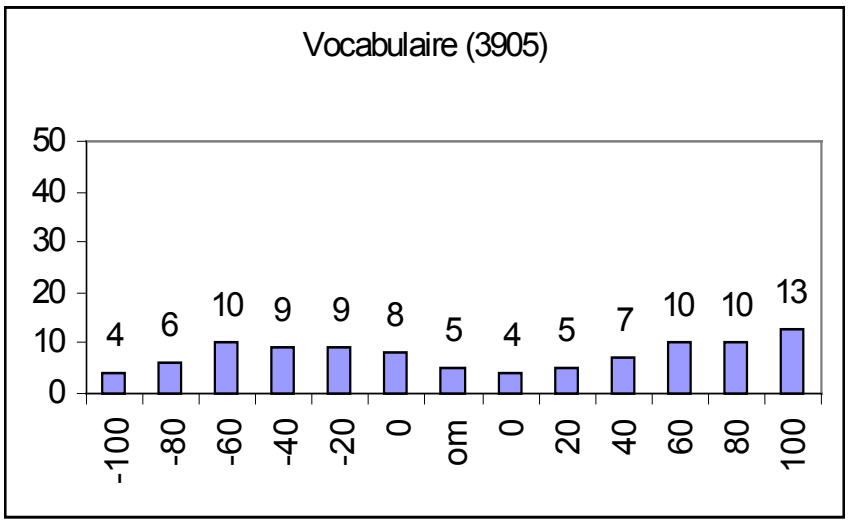
Ces données ont permis de donner des feedbacks cognitifs et métacognitifs à chaque étudiant, notamment en situant chaque réponse sur le spectre possible des performances allant de -100 à +100. Ces deux extrêmes représentent respectivement les réponses correctes avec certitude 100% (réponses parfaites) et les réponses incorrectes avec 100% (réponses dangereuses). Entre les deux, toute la gamme du pire au meilleur.

La distribution spectrale est un histogramme regroupant pour un étudiant, voire un groupe d'étudiants, la répartition des réponses à une épreuve sur ce continuum spectral.

Une courbe peut être tracée pour chaque hémispectre ; idéalement elle doit avoir une forme de J (la forme visée par les pédagogues). Son « asymétrie » et son escarpement peuvent être exprimés par des indices mathématiques, respectivement de Skewness et de Kurtosis.

Voici la répartition spectrale des réponses de 3905 étudiants ayant répondu aux 45 questions de l'épreuve de Vocabulaire dans l'opération MOHICAN. Les pourcentages (arrondis à l'unité) ci-dessous se rapportent donc à 175725 questions.

[1] Ce projet, signifiant Monitoring Historique des CANdidatures, a été mené par les toutes les universités de la CFWB. Dix épreuves (math, chimie, physique, bio, Vocabulaire, Syntaxe, Compréhension de Textes, Lecture de graphiques et de cartes Géo, Histoire-Actualité-Economie, Connaissances artistiques. Ce testing a touché près de 4000 étudiants de 8 des 9 universités, dans toutes les facultés et sections.



On voit que l'hémispectre de gauche a une forme gaussienne, mais que celui de droite a une forme tendant vers la courbe en J.

On peut aussi calculer pour chaque étudiant un score de « calibration » (souvent appelé Réalisme), dont le principe est de sommer les écarts entre les prédictions et la réalité, sur un graphique, les écarts entre les taux d'exactitude observés (en ordonnée) et les taux prédits, ou Degrés de Certitude (en abscisse), bref les écarts verticaux à la diagonale.

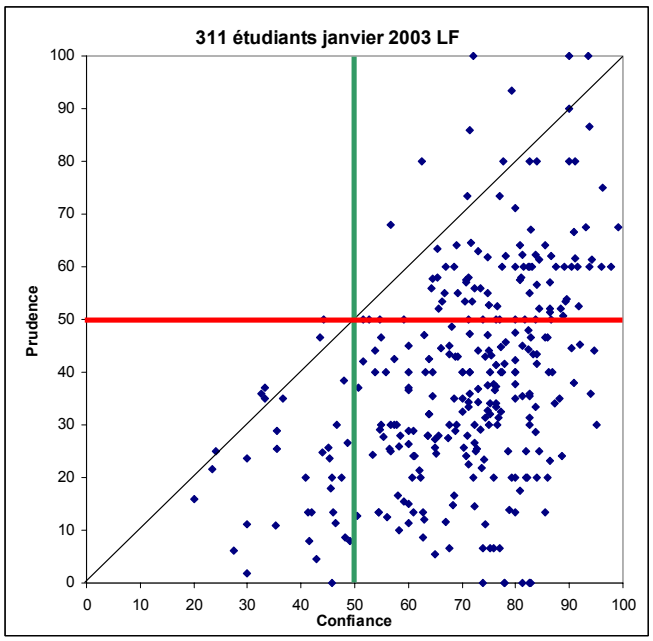
Comme de tels indices sont difficiles à calculer et surtout à interpréter, nous en avons proposé trois nouveaux qui se veulent faciles à calculer, diagnostics (précisant ce qui doit être amélioré) et faciles à utiliser (notamment dans des comparaisons intra étudiant ou inter-étudiants ou intra et inter-questions).

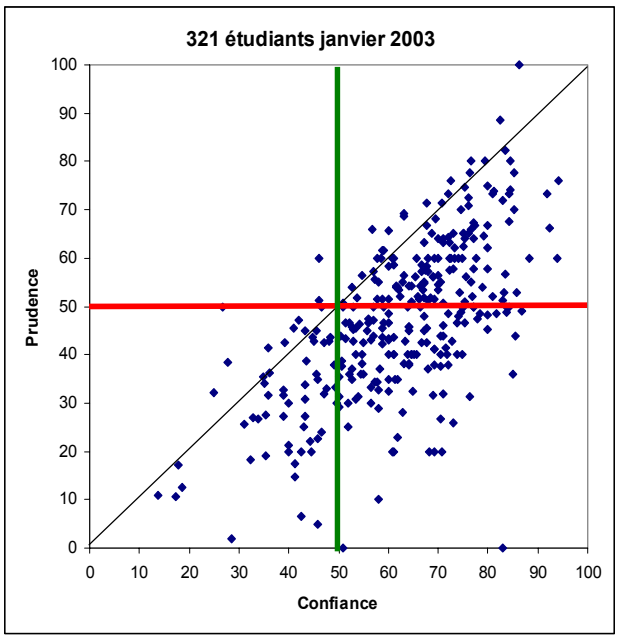
Il s'agit des indices de

- **Confiance** : la moyenne des degrés de certitude accompagnant les réponses **correctes.** Idéalement, elle devrait être la plus élevée possible (proche de 100%).
- **Prudence** : la moyenne des degrés de certitude accompagnant les réponses **incorrectes.** Idéalement, elle devrait être la plus faible possible (proche de 0%).
- **Discriminance (ou Nuance)** : la différence entre ces deux moyennes.

4. UNE OPÉRATION D'ÉVALUATION SANCTIONNANTE

Voici les indices de Confiance (en abscisse) et de Prudence (en ordonnée) pour un peu plus de 300 étudiants ayant passé deux interrogations dispensatoires, l'une (40 questions) à Livre Fermé, l'autre (40 questions) à Livre Ouvert. Chaque point représente un étudiant.





Les moyennes sont les suivantes :

	Livre Fermé	Livre Ouvert
Confiance	70%	62%
Prudence	38%	47%
Discriminance	32%	15%

On constate que la capacité des étudiants de discriminer est largement supérieure dans l'épreuve à livres fermés (32% de Discriminance moyenne) par rapport à l'épreuve à Livre Ouvert (15% seulement), et ce à la fois par une confiance plus élevée (70% au lieu de 62%) et une Prudence meilleure (38% au lieu de 47%).

Cette observation est à rapprocher de celles que nous avons faites dans la comparaison entre épreuves MOHICAN en partant des indices de Confiance, Prudence, Discriminance calculés question par question : « Dans les deux épreuves de connaissance de faits précis (en Art et en Histoire-Economie-Actualité), ...les manques de confiance et les manques de prudence sont minoritaires…Grosso modo, quand on sait, on sait qu'on le sait et quand on ignore, on en est conscient aussi. ».

Par contre,
« Dans l'épreuve de vocabulaire général de la langue française, qui combine la connaissance et la compréhension de concepts, on observe beaucoup plus d'imprudences (deux fois plus nombreuses que les manques de confiance).…Dans les épreuves de Compréhension (de textes et de graphiques) et de Syntaxe, les situations d'imprudence sont massives (Les indices de prudence moyens valent respectivement 55% et 60%). » (Leclercq et Poumay, 2003, 186).

Cette sensibilité des indices de Confiance, Prudence et Discriminance à la nature des questions est précieuse dans l'étude de la métacognition.

Au cours de l'année académique 2002-2003, nous avons instauré la notation certificative dans 5 cours universitaires sur base du barème de notation suivant :

> En cas de réponse correcte : + 1 ;
> en cas d'omission = 0 ;
> en cas de réponse incorrecte = -0,5.
> Le score total est ensuite ramené sur 20 points.
> L'étudiant reçoit ensuite des « **plus métacognitifs** » suivants :
> + 1 point si son indice de Confiance est supérieur à 50%,
> +1 point si son indice de Prudence est inférieur à 50%
> +1 point si son indice de discriminance est supérieur 20%.

Voici, pour l'épreuve de janvier à Livres Fermés, les trois indices de l'étudiant au matricule 587, ses « plus métacognitifs », son score classique et, finalement son score après l'ajout des « Plus métacognitifs » :

	Conf	Prud	Disc	C	P	D	Sc. class	SC. + Méta
587	58	40	18	1	1	0	7,9	9,9

Voici les mêmes indices pour une dizaine d'étudiants pour la même épreuve :

	Conf	Prud	Disc	C	P	D	Sc. class	SC. + Méta
597	64	53	10	1	0	0	15,3	16,3
1509	65	40	25	1	1	1	10,5	13,5
1557	73	43	30	1	1	1	8,9	11,9
1864	54	47	7	1	1	0	15,3	17,3
10503	75	55	20	1	0	0	13,7	14,7
10886	49	8	42	0	1	1	7,4	9,4
10998	64	47	18	1	1	0	4,7	6,7
11467	72	46	26	1	1	1	8,9	11,9
11490	75	57	19	1	0	0	10,5	11,5
11638	60	20	40	1	1	1	5,8	8,8
11660	70	23	47	1	1	1	12,1	15,1

En moyenne, pour 300 étudiants environ, le score est passé de 9,3 à 11,3. On voit que dans ce système de notation, l'utilisation par l'étudiant des degrés de

certitude ne peut lui être que favorable. En outre, la mesure de la connaissance et du réalisme sont totalement indépendants.

5. CONCLUSIONS

Notre conviction est que nous avons enfin atteint un degré de simplicité des consignes et des barèmes de scoring et une diagnosticité des indices telle que l'on va enfin pouvoir travailler sur l'aspect le plus intéressant de la métacognition : la réflexion du sujet sur ses propres performances, et la régulation de ses stratégies de réponses et d'apprentissage. C'est en cela que nous pensons que nos propositions constituent une démarche se distinguant par sa haute validité conséquentielle.

6. RÉFÉRENCES

Damasio, A. (2001). *L'erreur de Descartes.* Paris : Odile Jacobs.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, *18*, 87-123.

Jacobs, S.S. (1971). Correlates of unwarranted confidence in response to objective test items. *Journal of Educational Measurement*, *8,* 1.

Leclercq, D (1982), Confidence Marking : Its Use in Testing, in *Evaluation in Education*, vol. *6, 2*, 161-287.

Leclercq, D. (Ed.) (2003), *Diagnostic cognitif et métacognitif au seuil de l'université*. Le projet MOHICAN mené par les 9 universités de la Communauté française Wallonie Bruxelles. Liège : Editions de l'Université de Liège.

Leclercq, D., & Poumay, M. (2003) Analyses édumétriques et indices métacognitifs appliqués aux questions des 10 check-up MOHICAN, In D. Leclercq (Ed.) *Diagnostic cognitif et métacognitif au seuil de l'université*, Liège : les Editions de l'université de Liège, 181-190.

Sandbergen, S. (1968). Test strategie/test strategy. *Ned. T. Psychol.,* 23, 16-38.

Sandbergen, S. (1972). Guessing and confidence in testing educational achievement. *In* Choppin, B. *(A/106 IEA memorandum*).

Stevens, S.S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153-181.